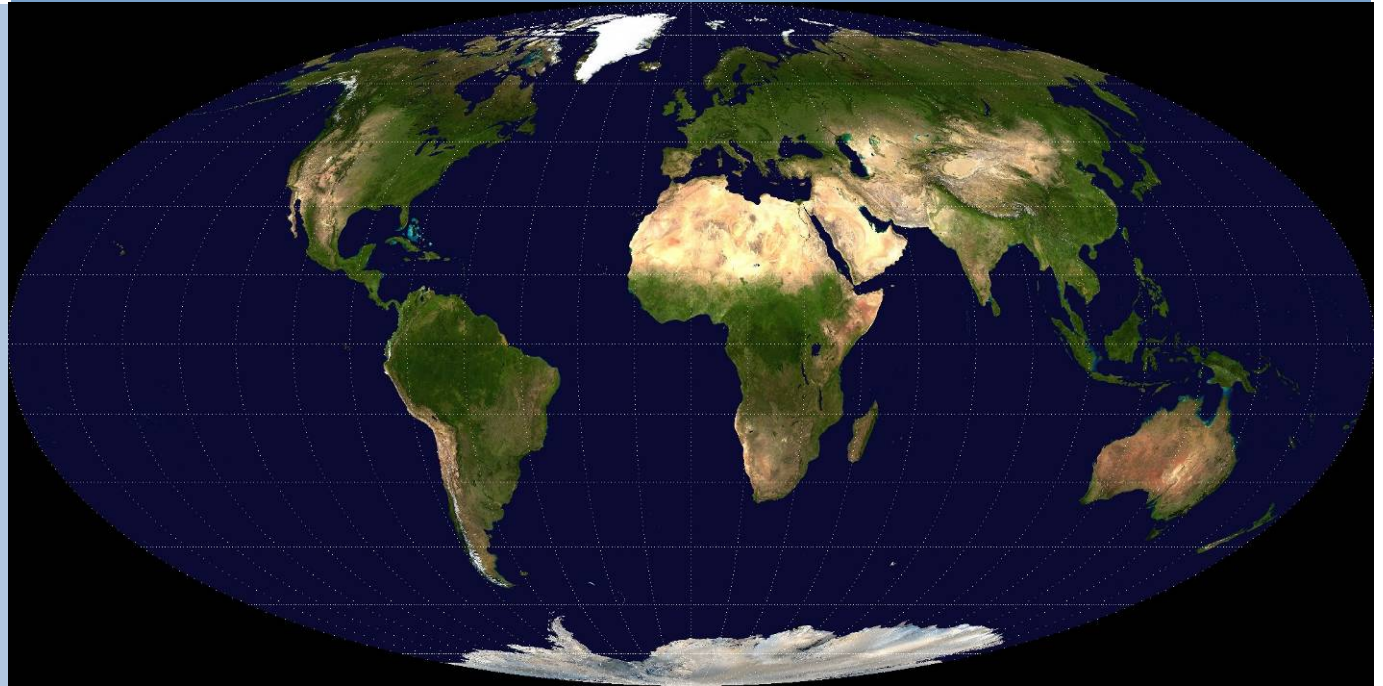


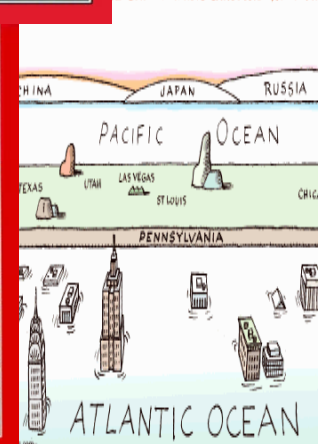
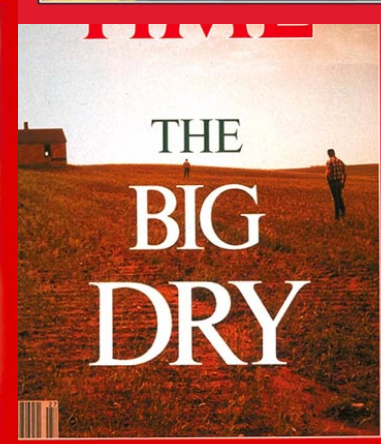
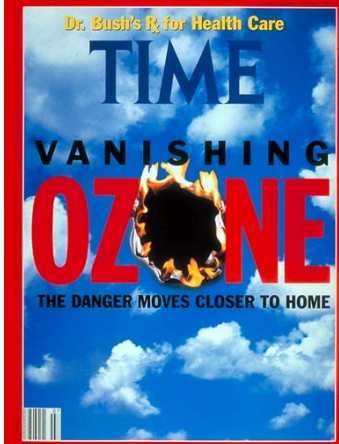
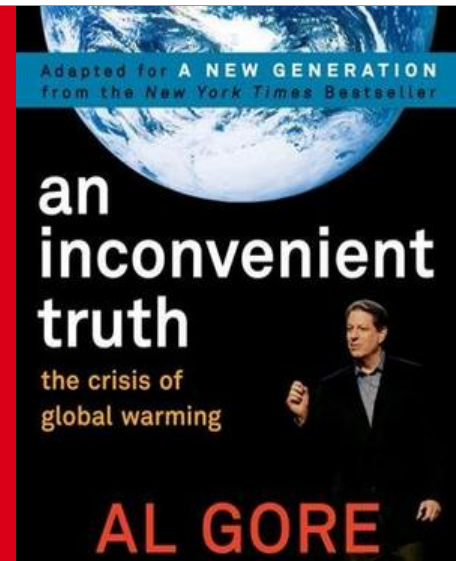
- Introductions
- Review/Intro to DataONE
- Summer Internship Program
 - Program and Objectives
 - Expectations
 - Resources
 - Status of projects
 - Steps forward

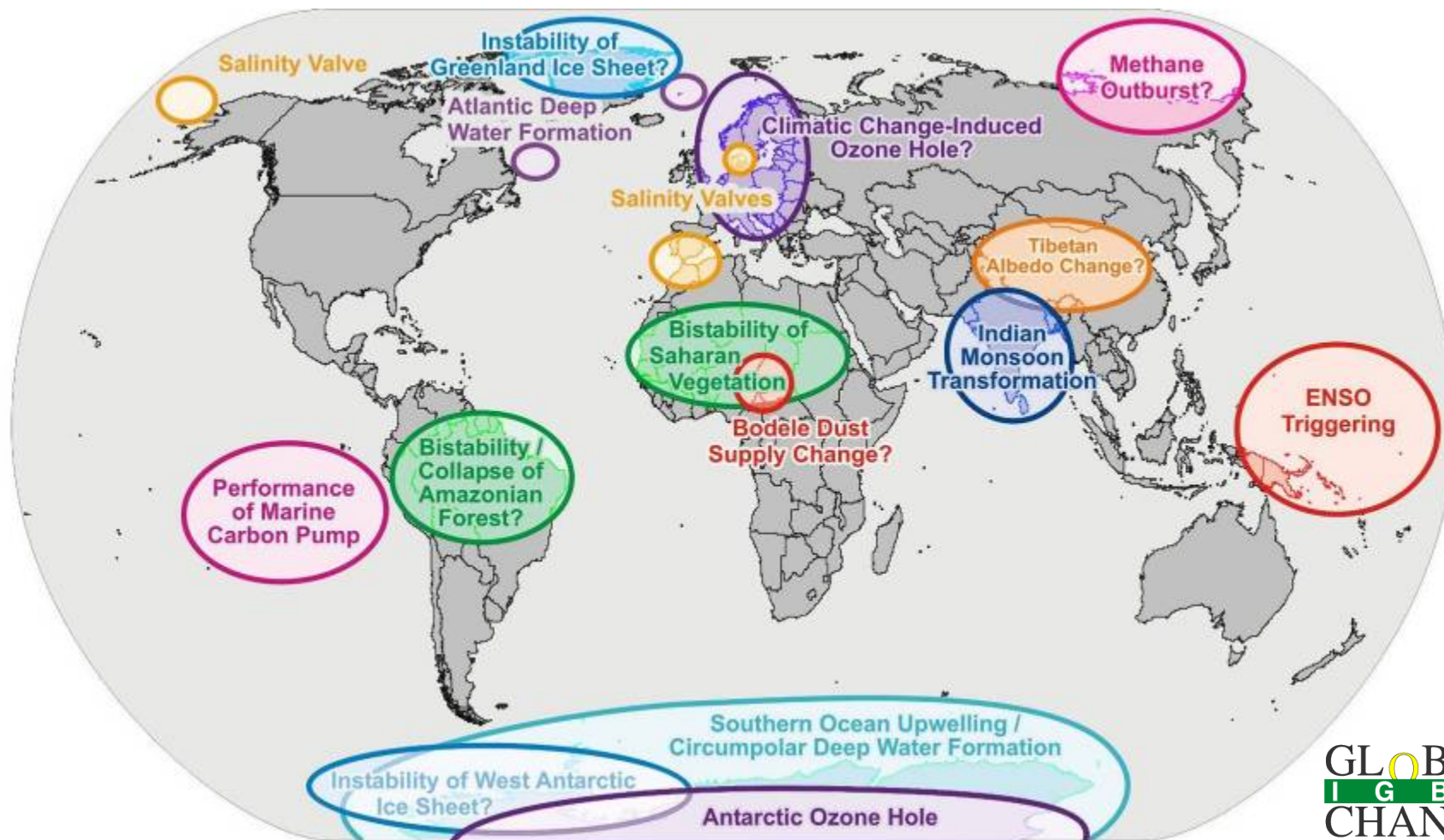


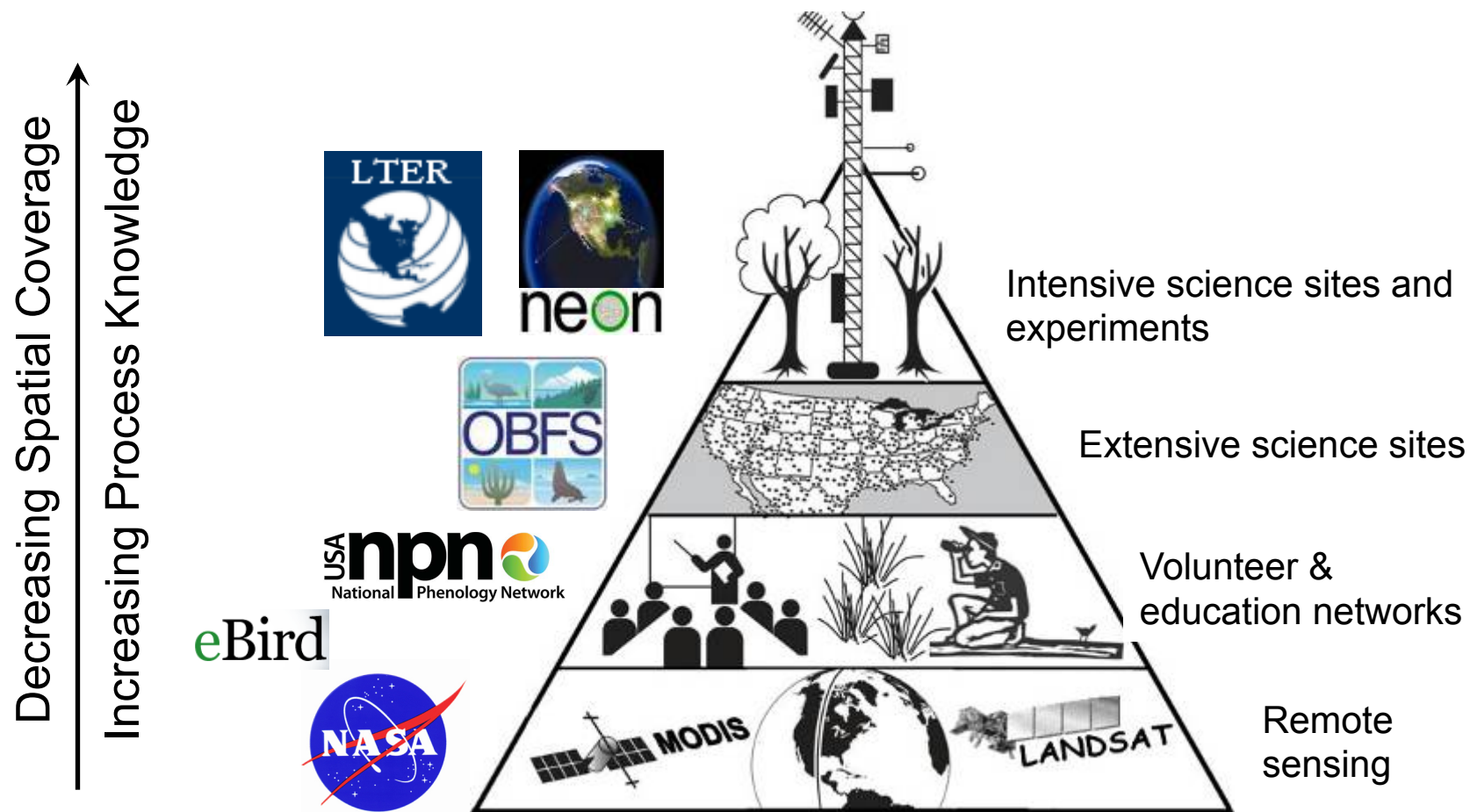
Building a virtual data center for the biological, ecological and environmental sciences

- William Michener (DataONE and University Libraries, University of New Mexico)
- Rebecca Koskela (DataONE, University of New Mexico)
- Dave Vieglais (DataONE and Biodiversity Research Center, University of Kansas)
- DataONE Team









Adapted from CENR-OSTP

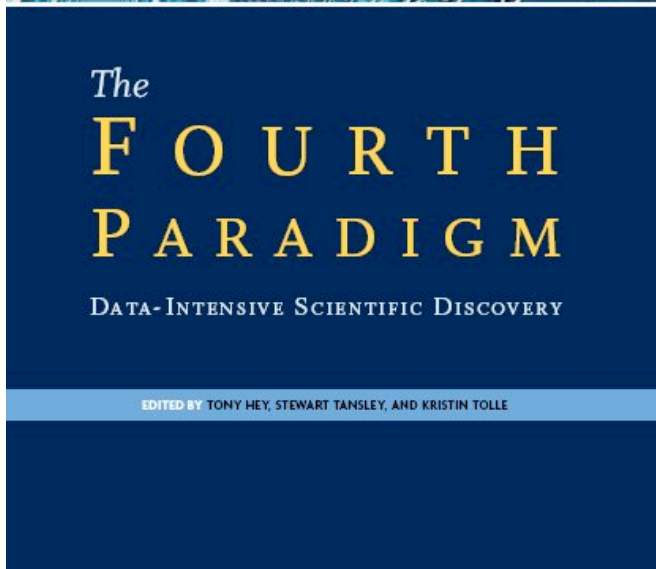
The Fourth Paradigm: “Data Intensive Research”

1. Experimentation → 2. Theory →
3. Computer simulations → 4. Data Intensive Research



“The impact of Jim Gray’s thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science.”

— Bill Gates, Chairman, Microsoft Corporation



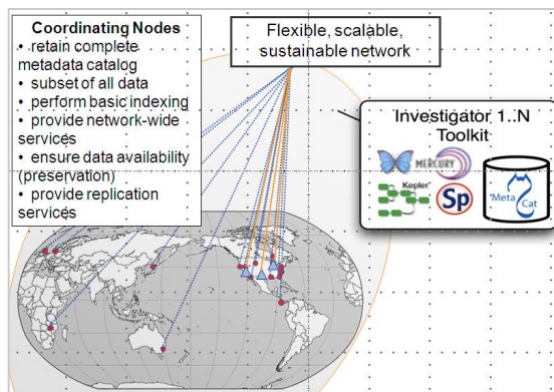
“One of the greatest challenges for 21st-century science is how we respond to this new era of data-intensive science. This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working.”

— Douglas Kell, University of Manchester

Providing *universal access to data about life on earth and the environment that sustains it*

- engaging the scientist in the data curation process
- supporting the full data life cycle
- encouraging data stewardship and sharing
- promoting best practices
- engaging citizens
- developing domain-agnostic solutions

1. Build on existing cyberinfrastructure



2. Create new cyberinfrastructure



3. Support new communities of practice

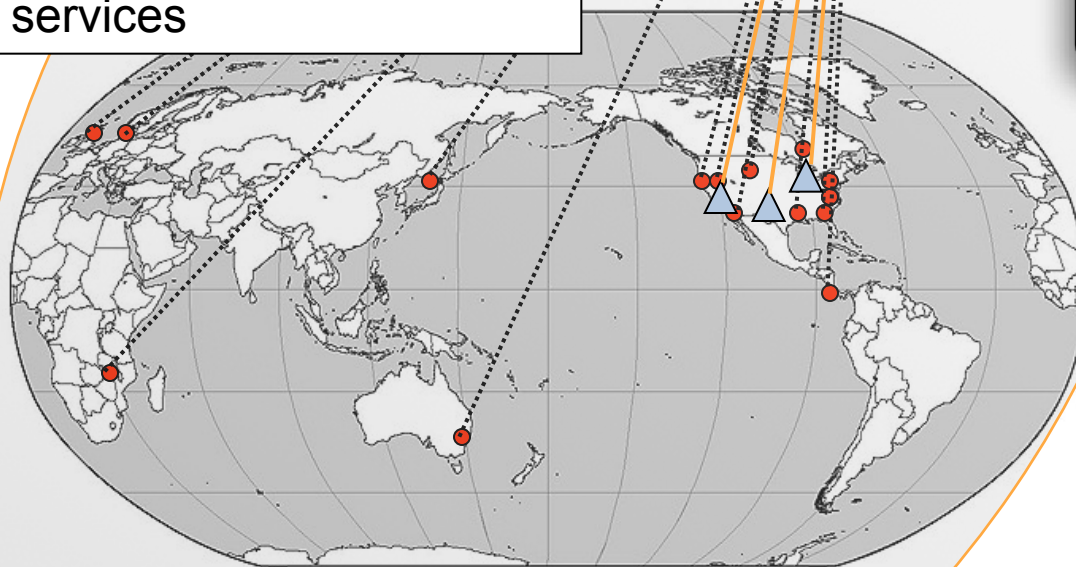
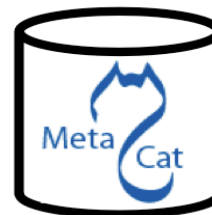
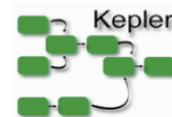


Coordinating Nodes

- retain complete metadata catalog
- subset of all data
- perform basic indexing
- provide network-wide services
- ensure data availability (preservation)
- provide replication services

Flexible, scalable,
sustainable network

Investigator 1..N Toolkit





A Pilot Catalog For
Earth Observations

DataONE Metadata Clearinghouse
HELP

Simple Search
Advanced Search

Search All Records For

Hint: boolean operators, wildcards and phrases are allowed.
ex: precipitation or (rain* and "moisture content")

Results/Page
10

Query being built:

Not Editable

1. Engage the community

- Perform baseline and iterative community assessments
- Usability studies
- DataONE Users Group



2. Leverage existing cyberinfrastructure

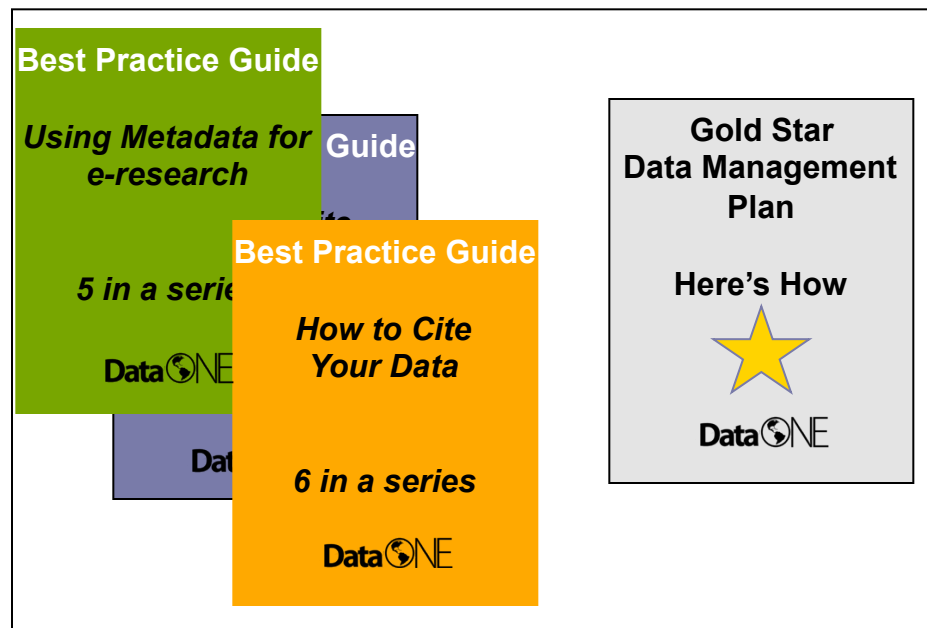


KNB, ESDIS, and Waters
Networks

3. Educate

Career Long Learning:


- best practice guides
- exemplary data management plans
 - podcasts, web-casts
- workshops and seminars
- downloadable curricula



4. Enable new science and demonstrate success

Pilot Catalog: Simple Search Interface

(searches entire metadata record)



The screenshot shows the DataONE Pilot Catalog Simple Search Interface. At the top, the DataONE logo is on the left, and the text "A Pilot Catalog For Earth Observations" is on the right. Below this is a navigation bar with "DataONE Metadata Clearinghouse" and a "HELP" link. The main content area has two tabs: "Simple Search" (selected) and "Advanced Search". Below the tabs, the text "Search All Records For" is centered. A search input field is followed by a "SEARCH" button. Below the input field, a hint reads: "Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain* and 'moisture content')". To the right of the input field is a "Results/Page" dropdown menu set to "10". Below the search area, the text "Query being built:" is followed by a large, multi-line text area. Below this text area, the text "Not Editable" is displayed. At the bottom left, there is a "CLEAR QUERY" link.

>70,000 Data Products

NBII Metadata Clearinghouse

Long Term Ecological Research (LTER) Network

ORNL Distributed Active Archive Center for Biogeochemical Data

Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA)

Organization of Biological Field Stations

Inter-American Institute for Global Change Research (IAI)

MODIS and ASTER Products (LPDAAC)

National Phenology Network (USANPN)

- University of New Mexico – William Michener, Rebecca Koskela, Mark Servilla
 - Cornell University – Paul Allen, Steve Kelling
 - CSIRO (Australia) – Donald Hobern
 - Duke University – Ryan Scherle, Todd Vision
 - Ecological Society of America – Cliff Duke
 - Oak Ridge National Laboratory – Bob Cook, John Cobb, Line Pouchard, Bruce Wilson
 - University of California Curation Center – Patricia Cruse, John Kunze
 - University of California Davis – Bertram Ludaescher
 - University of California Santa Barbara – Stephanie Hampton, Matt Jones
 - University of Edinburgh (Scotland) – Peter Buneman
 - University of Illinois – Urbana Champaign – Randy Butler, Von Welch
 - University of Illinois – Chicago – Robert Sandusky
 - University of Kansas – Dave Viegla
 - University of Manchester (UK) – Carole Goble
 - University of Michigan – Peter Honeyman
 - University of Southampton (UK) – David DeRoure
 - University of Southern California – Ewa Deelman
 - University of Tennessee – Suzie Allard, Carol Tenopir, Maribeth Manoff
 - USGS (National Biological Information Infrastructure, National Phenology Network) – Mike Frame, Vivian Hutchison, Jake Weltzin
 - Utah State University – Jeffery Horsburgh
 - EVA: Steve Kelling¹, Daniel Fink¹, Suresh Santhana Vannan², Kevin Webb¹, Robert Cook², William Michener³, Jeff Morissette⁴, Claudio Silva⁵, Tom Dietterich⁶, Patrick O’Leary⁷, Alyssa Rosemartin⁴, and Damian Gessler⁸
 - ¹Lab of Ornithology, Cornell University and ²Oak Ridge National Laboratory, ³University of New Mexico, ⁴U.S. Geological Survey, ⁵University of Utah, ⁶Oregon State University, ⁷Idaho National Laboratory, ⁸I-Plant
- NSF CISE Pathways Computational Sustainability and INTEROP, Leon Levy Foundation, and National Aeronautics and Space Administration

- Started in 2009 – 4 students
- 8 students selected for 2010
- 4 projects, students mostly working in teams
 - Multiple mentors
 - 1 project with 1 student
 - Some early start and late start
 - Incorporating Information Science for the first time
- Motivation
 - Engage students
 - Help develop new science data workers
 - Get useful work done in a learning environment

- Weekly (at least) discussion with students & mentors
- Progress reports posted
- Mid-term feedback (survey)
- Products posted to DataONE sites
 - Code in SubVersion (<https://repository.dataone.org>)
 - Documents posted in Plone site (<https://dataone.org>)

- <https://dataone.org/member-area/project-information/collaboration-tools>
- SubVersion repository (revision control)
- Plone site (for documents)
 - <https://dataone.org/member-area/projects/summer-internships/2010-intern-projects>
- Mailing lists and private e-mail
 - mentors@dataone.org, developers@dataone.org, Project specific lists
- Internet Relay Chat (IRC)
 - Mostly for developers: #dataone on irc.ecoinformatics.org
 - Web client available: <http://irc.ecoinformatics.org>
- Marratech (meet2.nceas.ucsb.edu)
- Etherpad (epad.dataone.org – for real-time notes)
 - Using

- Goal: Design and implement a "deep provenance store" (DPS) that combines provenance traces from the execution of different workflows, and that can be queried to explore lineage relationships across multiple workflow runs.
- Mentors:
 - Bertram Ludaescher (UC, Davis)
 - Paolo Missier (Manchester Univ)
 - Shawn Bowers (Gonzaga University)
- Students:
 - Anand Sarkar (UC, Davis)
 - Biva Shrestha (Appalachian State University)

- Goal: Implement a Member Node stack with Fedora Commons as a back-end data store
- Mentors:
 - Jerry Pan (ORNL)
 - Bruce Wilson (ORNL/UT)
 - Giri Palanisamy (ORNL)
- Student:
 - Makarand Bhonsle (Missouri U. Science & Tech)

- Goal: Analysis of citation and data reference practices, a review of existing materials on barriers to data sharing and reuse, an initial draft of a paper suitable for publication in a peer-reviewed journal analyzing sharing and reuse barriers, and informational materials suitable for use by scientists to describe best practices (including a listing of exemplar data management plans)
- Mentors:
 - Suzie Allard (University of Tennessee)
 - Maribeth Manoff (University of Tennessee)
 - Todd Vision (NESCent/U North Carolina)
 - Bruce Wilson (ORNL/UT)
- Students:
 - Valerie Enriquez (Simmons University)
 - Nicholas Weber (University of Illinois)
 - Sarah Judson (Brigham Young University)

- Goal: Survey the attitudes of academic and government librarians to data sharing, including perceptions of benefits and barriers.
- Mentors:
 - Carol Tenopir (University of Tennessee)
 - Suzie Allard (University of Tennessee)
 - Robert Sandusky (University of Illinois, Chicago)
- Students:
 - Christine Murray (University of Michigan)
 - Elizabeth Allen (University of Illinois)



Path Forward

